JOSCEX
Journal of Soft Computing Exploration

SVM Optimization with Correlation Feature Selection Based Binary Particle Swarm Optimization for Diagnosis of Chronic Kidney Disease

Doni Aprilianto
Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Jul 24, 2020
Revised Aug 6, 2020
Accepted Sept 2, 2020

Keywords:

Support Vector Machine;
Correlation-Based Feature Selection;
Chronic Kidney Disease;
Binary Particle Swarm Optimization.;

ABSTRACT

Data mining has been widely used to diagnose diseases from medical data. In this study using chronic kidney disease dataset taken from UCI Machine Learning. The dataset has 25 attributes with 400 samples. With 25 attributes that allow redundant data. Redundant data in datasets can reduce computational efficiency and classification accuracy. To increase accuracy of classification algorithm can be done by reducing dimensions of dataset. Correlation-based Feature Selection (CFS) can quickly identify and filter redundant attributes. However, CFS has disadvantage that selected attribute is not necessarily the best attribute. These weaknesses can be overcome by Binary Particle Swarm Optimization (BPSO). BPSO chooses attributes based on the best fitness value. The purpose of this study is to improve accuracy of Support Vector Machine (SVM) by implementing combination of CFS and BPSO as feature selection. Accuracy of SVM in predicting CKD is 63.75%. Whereas, accuracy of SVM by applying CFS as feature selection is 88.75% and average accuracy of ten execution SVM algorithms by applying a combination of CFS and BPSO as feature selection is 95%. Thus, combination of CFS and BPSO as feature selection on the SVM algorithm can improve results of accuracy in diagnosing CKD by 31.25%.

Corresponding Author:

Don Aprilianto
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: doniapr14@students.unnes.ac.id

1. INTRODUCTION

In recent years, data mining has been widely used in the health sector, bioinformatics, banking, document classification marketing, etc. [1]. in the health sector, data mining is used to diagnose diseases such as breast cancer, diabetes, heart disease and others [2]. To predict a decision in data mining can use classification techniques. Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Artificial Neural Network are examples of classification algorithms [3].

Popular algorithm for data classification is support vector machine (SVM). However, SVM requires a large amount of memory if the processed data has high dimensions or has a lot of features [4]. Accuracy and computational efficiency of a classification algorithm are strongly influenced by datasets that process, datasets with high dimensions can present data that is redundant and irrelevant so that it will affect the performance of the classification algorithm [5].

To improve accuracy and performance of classification algorithm can be done by reducing dimensions and eliminating redundant data. This process is called feature selection. The feature selection algorithm is divided into 3 main categories namely, filter, wrapper and hybrid. The filter method has a fast computation time, but not necessarily finding the best combination of subset attributes. The wrapper method will produce the best combination of subset attributes, but requires large computer memory. The hybrid method will combine the filter method and the wrapper method to get the advantages of each method. In hybrid method, filter method is used to reduce dataset dimensions and wrapper method will be used to find the best combination of attributes [6].

One of the algorithms in filter method that effective for handling redundant and irrelevant data is correlation-based feature selection (CFS). CFS is an algorithm that ranks subset attributes and finds relevant attributes based on correlation-based heuristic evaluation functions [7]. CFS can quickly identify and filter out irrelevant and redundant attributes [8]. CFS will choose attributes that have a strong correlation with the target class but do not correlate with other attributes. However, CFS does not necessarily choose attributes that provide the best accuracy results if the data sample is limited [5]. To get the best combination of attributes and good correlation, CFS can be combined with the wrapper method.

Over the past few years, many wrapper methods have been developed for attribute selection. An example of a wrapper algorithm developed is evolutionary programming (EP), ant colony optimization (ACO), differential evolution (DE), genetic algorithms, particle swarm optimization (PSO) [9]. For the optimization, PSO gives more competitive results than Genetic Algorithms [2]. At PSO, each particle is flown in search space to find the best solution (fitness) called pbest. Then, the overall best value is called gbest. To overcome the feature selection problem, the particles in PSO will be represented in binary form which is then called Binary Particle Swarm Optimization [10].

In this study propose a combination of CFS algorithm and BPSO algorithm as feature selection to improve the accuracy of the SVM algorithm for diagnosing chronic kidney disease (CKD). CFS was chosen because it can choose attributes that have good correlation results and delete redundant data quickly, while BPSO was chosen because it was able to provide a combination of attributes that produced the best accuracy.

2. METHOD

In this study, the combination of CFS and BPSO was carried out as a feature selection. CFS is used to reduce the dimensions of the dataset based on the correlation between features and target class but does not correlate with other features. BPSO is used to find the best combination of features. The classification method used is the Support Vector Machine algorithm. From the classification results, we will get an increase in accuracy from Support Vector Machine before and after the combination of CFS and BPSO is applied. The flowchart of the method used in this study is shown in Figure 1.

2.1 Data preprocessing

The data used in this study is the Indian Chronic Kidney Disease taken from the UCI Machine Learning Repository. This dataset has 25 attributes, of which 11 attributes are numeric and 14 are nominal.

The CKD dataset has 400 data samples and there are more than 15% missing values. With a missing value of more than 15% it will greatly affect the performance of the classification model that has been formed [11]. In this study, the handling of the missing value is done by using the most frequent.

2.1.1 Correlation-based feature selection

Correlation-based Feature Selection (CFS) is a filter algorithm that ranks subset attributes according to heuristic evaluation functions based on correlation [12]. CFS will evaluate features by considering the predictive capabilities of each feature and the level of redundancy between them. If the correlation between attributes and class is known, and the correlation between each attribute is given, then the correlation can be predicted by using Eq. 1.

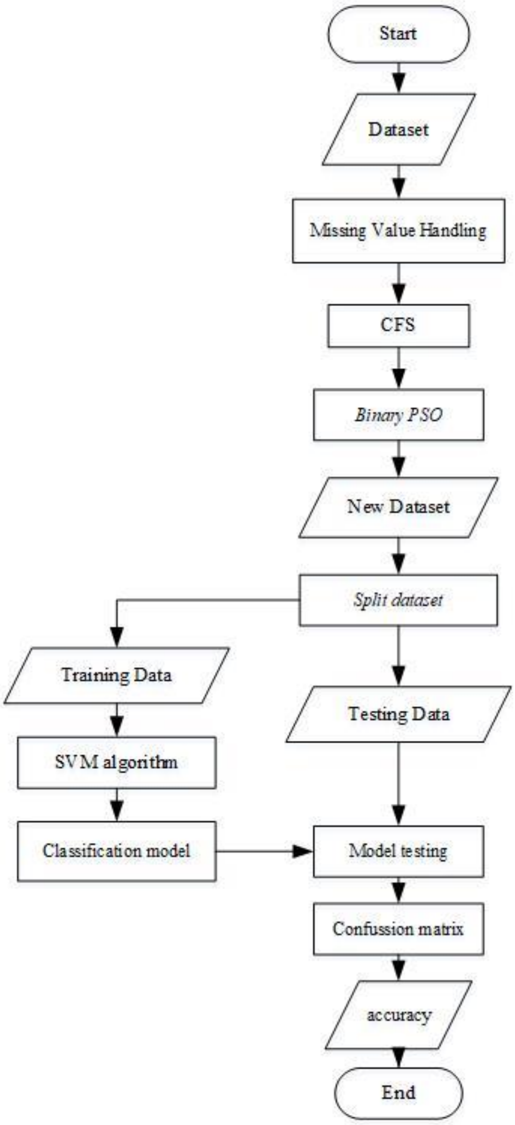

Figure 1. Flowchart SVM with CFS and BPSO as feature selection.

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \tag{1}$$

where Ms is correlation between summed components and external variables. K is number of components. rcf is average correlation between components and outside variables.rff is correlation between average components.

CFS is an automatic algorithm that does not require users to specify the number of features to be selected. CFS uses the best first search method to get the best features from a feature subset. Then the correlation between features can be calculated using symmetrical uncertainty (SU) as Eq. 2.

$$SU = 2.0 \times \left[\frac{H(S_j) + H(S_i) - H(S_i,S_j)}{H(S_j) + H(S_i)}\right] \tag{2}$$

Where H(Sj) dan H(Si) is entropy of attributes. H(S_i, S_j) is conditional entropy.

2.1.2 Binary particle swarm optimization

Binary Particle Swarm Optimization (BPSO) was introduced by Kennedy and Eberhart in 1997. BPSO is used to solve discrete optimization problems. The difference between PSO and BPSO lies in the representation of the particles [10]. In BPSO each particle is shown in discrete values. While in standard PSO, particles are represented in continuous values. The concept of PSO is that each particle is flown in search space to find the best solution (fitness) called pbest. Then, the best overall value (global value) called gbest. Each particle has two vectors namely position vectors and velocity vectors to move around in search space. Each particle has memory and each particle will track the best position beforehand [13].

The stages of the binary particle swarm optimization algorithm in the feature selection are as follows:

Step 1 : Initialize particles with random velocity and positions.
Step 2 : Calculate the fitness value of each particle in the population.
Step 3 : If fitness value of particle i is smaller than the fitness value of pbest, then set pbest from particle i to particle position i.
Step 4 : If pbest value is less than the current gbest value, then set gbest to the current pbest.
Step 5 : Update position and velocity of the particles by using Eq. 3 and 4.

$$Vid_{k+1} = w \times \ Vid_k + c_1 \times rand1 \times (Pid - Xid) + c_2 \times rand2 \times (Gid - Xid) \tag{3}$$

$$Xid_{k+1} = Xid_k + Vid_{k+1} \tag{4}$$

Step 6 : Where Vid is the individual velocity. Xid is position of individual. w is inertia weight parameter. c_1 and c_2 is learning rate constant, the value is between 0 and 1. rand1 and rand2 is random parameter between 0 and 1. Pid is Pbest (personal best) individual i in d dimension. Gid is Gbest (global best) in d dimensions.
Step 7: Iteration will stop if maximum generation is fulfilled; if not return to step 2.

2.2 Support vector machine

Support Vector Machine (SVM) was first proposed by Vladimir Vapnik. Proposed in the field of statistical learning theory and structural risk minimization [14]. SVM has been used in a variety of problems such as data classification, image classification, text categorization, tone recognition, digit recognition of handwriting [15].

The stages of the Support Vector Machine algorithm in classifying datasets are as follows:
Step 1: Prepare training data. The training data consists of 80% of the entire dataset.
Step 2: Finding boundaries between classes. When each point in a class is connected to another point, a line that separates between the classes will appear. This limit is known as the convex hull. Each class has its own convex hull and because the class (assumed) is linearly separated, this hull does not intersect.
Step 3: Determine a hyperplane that maximizes the margin between classes. Can be done in the following ways:

a. First, any hyperplane stated in two attributes, $x_1$and x_2, can be written Eq. 5.

$$w \cdot x + b = 0 \tag{5}$$

where w is weight $(w = w_1, w_2, \dots, w_n)$, x= number of attributes $(x = x_1, x_2, \dots, x_n)$, b = bias.

b. An optimal hyperplane, defined uniquely by $b_0 + w_0 \cdot x = 0$. After defining a hyperplane in this mode, it can determine the margin. Margin can be written Eq. 6.

$$margin = \frac{2}{\sqrt{w_0}} . w_0 \tag{6}$$

c. Maximizing this quantity requires quadratic programming, which is a process that has a strong position in the theory of mathematical optimization. Furthermore, w can be easily stated in some examples of training data, known as support vectors, can be written Eq. 7.

$$|w_0 = \sum y_i x_i| \tag{7}$$

where y_i is the class label and x_i is called support vector. i is a zero coefficient only for support vector.

Step 4: After setting boundaries and hyperplane, each new test can be classified by calculating on which side of the data results in the hyperplane. This can be found by replacing the test x example into the hyperplane equation. If you count +1, then it includes a positive class and if it is calculated as -1, then it belongs to the negative class.

2.3 Evaluation

The proposed method begins by dividing the dataset into training data and testing data. In this study the data distribution was done using the splitter method. This method divides the data into two subsets with a proportion of 80% for training data and 20% for testing data.

Measurement of classification performance is done by confusion matrix, confusion matrix is a useful tool to analyze how well the classifier recognizes tuples from different classes. Confusion matrix is done by calculating the number of predicted classes against the actual class. These results are expressed in True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). FP and FN state that the classifier is wrong in recognizing tuples, meaning positive tuples are recognized as negative and negative tuples are recognized as positive. While TP and TN state that the classifier recognizes tuples correctly, meaning positive tuples are recognized as positive and negative tuples are recognized as negative. The matrix confusion table can be shown in Table 2.

Table 1. Confusion matrix

Classification		**Predicted class results**		
		Yes	**No**	Total
Actual Class	**Yes**	TP	FN	P
	No	FP	TN	N
	Total	*P'*	*N'*	P+N

Accuracy is the percentage of the total data classified correctly. Accuracy measurements can be written by using Eq. 8.

$$Accuracy = \frac{TP+TN}{P+N} \text{ x } 100\% \tag{8}$$

3. RESULT AND DISCUSSION

In this study, the proposed algorithm testing uses the Python programming language by utilizing a scikit-learn library, sk-feature and pyswarms library. The data used is the CKD dataset taken from the UCI Machine Learning. This dataset has 25 attributes that have 1 class and 24 attributes.

3.1 Correlation Based Feature Selection Process

CFS will choose attributes that have the highest correlation weighting value. From the CFS process, 13 selected attributes were obtained. The list of attributes and weights of the CFS process is shown in Table 3.

Tabel 2. List of attributes and results of CFS weight

No	Attributes	CFS Weight
1	*Pc*	0,54636158
2	*Pe*	0,54568856
3	*Appet*	0,54506278
4	*Bp*	0,54405311
5	*Rc*	0,54397886
6	*Ane*	0,54383461
7	*rbc*	0,54211877
8	*Cad*	0,53743288

9	*Al*	0,53624068
10	*Pcv*	0,53390589
11	*Dm*	0,50241938
12	*Sg*	0,4529584
13	*Htn*	0,35495722

3.2 Binary Particle Swarm Optimization Process

Attributes chosen by the CFS algorithm, do not necessarily produce the best combination of attributes. Therefore, the BPSO algorithm is used to determine the best feature combination of the attributes chosen by CFS. At this stage, 10 tests are executed to determine the best combination of features. The BPSO parameters used in this study are shown in Table 4.

Table 3. BPSO parameters

Parameters	Value
Number of particles	60
Iteration	100
Inertia weight	0,9
C1	2
C2	2

3.3 Application of Algorithms

At this stage, 3 tests were carried out, namely stand alone SVM algorithm, SVM algorithm by implementing CFS and SVM algorithm by implementing a combination of CFS and BPSO. In the first application, the SVM algorithm will process the CKD dataset with 25 attributes. The application of SVM algorithms gets an accuracy of 63.75%. The results of this accuracy state that the SVM algorithm can classify CKD datasets well because the accuracy results are greater than the error rate. However, the results of this accuracy can be improved by applying several preprocessing methods.

The second application, the SVM algorithm will be combined with the CFS algorithm. So SVM will process the CKD dataset with 13 attributes and 1 class. The accuracy of this classification model is 88.75%. The accuracy of the application of this model can increase the accuracy of the SVM algorithm by 25%. However, these results can still be improved by selecting the best feature combination using the BPSO algorithm.

The third application, the SVM algorithm will be combined with the CFS and BPSO algorithms. In this implementation, 10 tests were executed to determine the best combination of features. The accuracy of this classification model can be seen in Table 5

Table 4. Average SVM Accuracy Results with a combination of CFS and BPSO as feature selection

Execution	Number of attributes	Accuracy (%)
1	12	96,25 %
2	10	95 %
3	11	95 %
4	10	96,25 %
5	10	93,75 %
6	11	95 %
7	11	96,25 %
8	12	96,25 %
9	9	91,25 %
10	10	95 %
mean	10,6	95

The accuracy of the application of this model can increase the accuracy of the SVM + CFS algorithm by 6.25% and can increase the SVM algorithm by 31.25%. comparison of accuracy of each application of the algorithm can be seen in Figure 2.

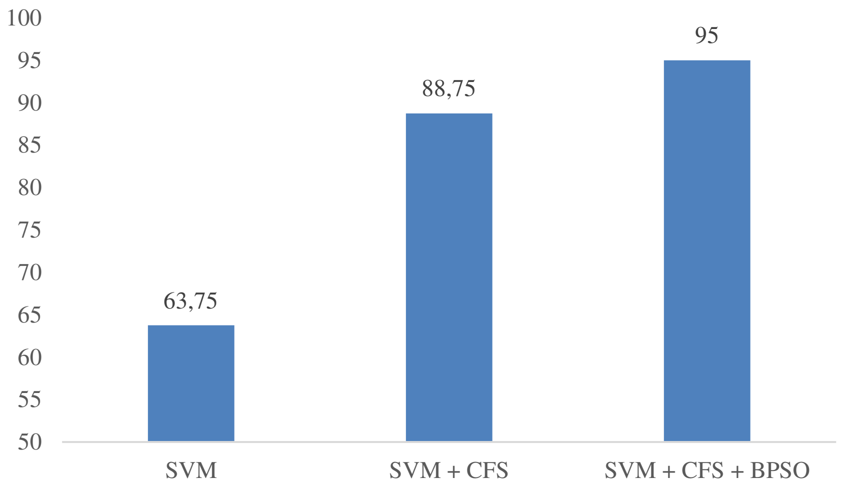

Figure 2. The comparison result of accuracy

4. CONCLUSION

In this study, the testing of SVM algorithm by applying the CFS algorithm and the BPSO algorithm was carried out using the CKD dataset taken from the UCI Machine Learning Repository. CFS algorithm is used to get attributes with good correlation while BPSO is used to get the best combination of attributes. The results of this study showed that the accuracy of the application of the SVM algorithm was 63.75%, while after the CFS algorithm was used, the accuracy of 88.75% and the accuracy of ten SVM algorithms was obtained by applying a combination of feature selection CFS and BPSO of 95 %. Thus, it can be concluded that the application of a combination of CFS and BPSO as feature selection on SVM algorithm can improve the results of accuracy in diagnosing CKD by 31.25%.

REFERENCES

[1] C. Sreedhar, N. Kasiviswanath, and P. C. Reddy, “Clustering large datasets using K-means modifed inter and intra clustering (KM-I2C) in Hadoop, ” *J. of Big Data*, vol. 27, no. 4, pp. 1-19, 2017.

[2] M.A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, “Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis, ” presented at the 5th Int. Conf on Mathematics, Science and Education, Bali, Indonesia, Oct. 8–9, 2018.

[3] D. O. Sahin, and E. Kılıc, “Two new feature selection metrics for text classification, ” *Automatika*, vol.60, no. 2, pp. 162-171, 2019

[4] V. Kotu, and B. Deshpande, Predictive Analytics and Data Mining. Massachusetts, USA: Morgan Kaufmann, 2015, pp. 63-163.

[5] I. Jain, V. K. Jain, and R. Jain, “Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification, ” *Appl. Soft Comp.,* vol. 62, no.-, pp. 203-215. 2018.

[6] K. Sutha, and J. J. Tamilselvi, “A review of feature selection algorithms for data mining techniques, ” *Inte. J. on Comp. Sci. & Eng.*, vol. 7, no. 6, pp. 63-67, 2015.

[7] P. Yildirim, “Filter based feature selection methods for prediction of risks in hepatitis disease, ” *Int. J. of Mach. Lear. & Comp.*, vol. 5, no. 4, pp. 258-263. 2016.

[8] S. Sasikala, S. Appavu, and S. Geetha, “Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set, ” *Appl. Comp. & Info.,* vol. 12. no.-, pp. 117-127. 2017.

[9] I.A. Ashari, M.A. Muslim, and Alamsyah. “Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing, ” *Sci. J. of Info.* vol. 3, no. 2, pp. 149-158, 2016.

[10] M.S. Muhammad, K.V. Selvan, S.M.W. Masra, Z. Ibrahim, and A.F.Z. Abidin, “An Improved Binary Particle Swarm Optimization Algorithm for DNA Encoding Enhancement, ” presented at the IEEE Symposium on Swarm Intelligence, Paris, France, April 11-14, 2011.

[11] W. Abedalkhader, and N. Abdulrahman, “Missing Data Classification of Chronic Kidney Disease, ”, *Int. J. of Data Mining & Knowledge Manage. Process*, vol. 7, no. 5, pp. 55-61.2017

[12] N. Gopika, and A. M. E. M. Kowshalaya, “Correlation based feature selection algorithm for machine learning, ” presented at the 3rd Int. Conf. on Commu. & Elec. Sys., Coimbatore, India, Oct. 15-18, 2018.

[13] N. D. Jana, and J. Sil, “Interleaving of Particle Swarm Optimization And Differential Evolution Algorithm For Global Optimization, ” *Int. J. of Comp. & Appl.,* vol. 38. no. -, pp. 116-133, 2016.

[14] J. Nayak, B. Naik, and H. S. Behera, “A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges, ” *Int. J. of Data. Theory & Appl.*, vol. 8, no.-, pp. 169-186. 2016.

[15] D. K. Srivastava, and L. Bhambhu, “Data classification using support vector machine, ” *J. of Theoretical & Appl. Inf. Tech*., vol. 12, no.-, pp. 1-7. Feb 2010.